Genetic Epidemiology 25: 136–148 (2003)

Sample Size Calculations for Population- and Family-Based Case-Control Association Studies on Marker Genotypes

Ruth M. Pfeiffer* and Mitchell H. Gail

Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland

Most previous sample size calculations for case-control studies to detect genetic associations with disease assumed that the disease gene locus is known, whereas, in fact, markers are used. We calculated sample sizes for unmatched case-control and sibling case-control studies to detect an association between a biallelic marker and a disease governed by a putative biallelic disease locus. Required sample sizes increase with increasing discrepancy between the marker and disease allele frequencies, and with less-than-maximal linkage disequilibrium between the marker and disease alleles. Qualitatively similar results were found for studies of parent offspring triads based on the transmission disequilibrium test (Abel and Müller-Myhsok, 1998, Am. J. Hum. Genet. 63:664–667; Tu and Whittemore, 1999, Am. J. Hum. Genet. 64:641–649). We also studied other factors affecting required sample size, including attributable risk for the disease allele, inheritance mechanism, disease prevalence, and for sibling case-control designs, extragenetic familial aggregation of disease and recombination. The large sample-size requirements represent a formidable challenge to studies of this type. *Genet Epidemiol* 25:136–148, 2003. Published 2003 Wiley-Liss, Inc.[†]

Key words: case-control study; family-based case-control study; trend test; correlated binary data; power calculations; linkage disequilibrium

Correspondence to: Ruth M. Pfeiffer, National Cancer Institute, 6120 Executive Blvd. EPS 8030, Bethesda, MD 20892-7244.
E-mail: pfeiffer@mail.nih.gov
Received for publication 18 December 2002; Revision accepted 17 February 2003
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.10245

INTRODUCTION

The proper use of association studies to identify disease-related genes is a topic of active research. Association studies based on marker alleles in linkage disequilibrium with a disease-producing allele have been proposed for fine-scale mapping in a region thought to contain a disease gene [Thompson et al., 1988; Olson and Wijsman, 1994; Kaplan and Morris, 2001]. We call such marker alleles ''markers'', to distinguish them from putative disease-causing alleles at disease loci, sometimes called ''candidate genes'', ''liability genes'', or ''disease genes''. Association studies based on markers have also been proposed as a means of screening the entire genome [Risch and Merikangas, 1996; Kruglyak, 1999], though serious questions have been raised about the feasibility of this approach for complex diseases [Weiss and Terwilliger, 2000].

Despite the substantial work that has been carried out to assess the power and sample size requirements for association tests based on markers, including work on the transmission disequilibrium test [TDT; Abel and Müller-Myhsok, 1998; Tu and Whittemore, 1999], there remains a gap in this literature for genotype-based tests in case-control studies with population-based unrelated controls and for family-based case-control studies. Lange and Laird [2002a] gave power calculations for a wide range of family-based designs based on genotypes of disease alleles, but not markers. Olson and Wijsman [1994] computed power for marker genotype-based tests with population-based case-control designs, in which it was assumed that the genotype of the disease gene could be inferred from the phenotypes of the cases and controls. This assumption would not apply to complex diseases. Kaplan and Morris [2001] investigated the power of association tests using a marker allele-based chi-square test statistic for unrelated cases and controls. This test statistic, as they pointed out, is not robust to departures from

Hardy-Weinberg equilibrium caused by inbreeding [see also Sasienei, 1997]. Schaid and Rowland [1998] discussed the power of score tests for association studies for candidate loci (but not markers) for case-control studies with several choices of family-based controls as well as independent controls. Slager and Schaid [2001] studied the power of the Armitage test for trend, based on the genotypes of a disease gene in unrelated cases and controls. The use of such two-sided trend tests was recommended for whole-genome scans based on biallelic markers such as single-nucleotide polymorphisms (SNPs) [Devlin and Roeder, 1999]. As this test is genotype-based, it is robust to allelic correlations induced by inbreeding, but not necessarily to population stratification or other types of confounding. We extend this work by calculating sample sizes and power to assess marker-disease associations for the trend test when the marker is in linkage disequilibrium with the disease gene. We also assess the power of genotype-based tests with markers when sibling controls, instead of population controls, are used. Such tests, though less efficient than tests based on unrelated cases and controls [Witte et al., 1999; Slager and Schaid, 2001], are robust to population stratification.

STATISTICAL METHODS

SAMPLE SIZES FOR INDEPENDENT CASES AND CONTROLS

Following the notation in Slager and Schaid [2001], we assume a random sample of R cases and S unrelated controls. Let the biallelic marker have genotypes *aa*, *aA*, and *AA*, and define a random variable $M = 0, 1, 2$, which corresponds to the numbers of A alleles in the marker genotype. The case-control data can then be summarized in a 2×3 table, where the columns correspond to genotype, M, and the rows to disease status, Y (see Table I).

Let $X(M = i) = X_i$ be a score associated with marker genotype $M = i$, for $i = 0, 1, 2$. If A were a disease gene, rather than a marker, the X_i would be the efficient scores for testing $\beta = 0$ in the model logit $P(Y = 1 \mid M = i) = \mu + \beta X_i$ [Armitage, 1955]. For example, the additive scores $X_i = i$, recommended by Devlin and Roeder [1999], would be efficient if the logit increased linearly with the number of alleles A. Dominant ($X_i = 1$ if $i = 1$ or 2, and 0 otherwise) or recessive ($X_i = 1$ if $i = 2$, and 0 otherwise) scores could be used as well. We use the scores that are efficient for disease genes but emphasize additive scores for markers for reasons explained later.

TABLE I. Genotype Counts for Cases and Controls

	M=0	M=1	M=2	
Cases ($Y = 1$)	r_0	r_1	r_2	R
Controls ($Y = 0$)	s_0	s_1	s_2	S
Total counts	n_0	n_1	n_2	N

The Armitage score statistic for $\beta = 0$ that tests for a trend in proportions is $U = X'[(1 - \phi)r - \phi s]$, where $\phi = R/N$ is the proportion of cases in the case-control study with $N = R + S$. The vector of scores is $X' = (X_0, X_1, X_2)$, and the genotype counts for cases and controls, $r' = (r_0, r_1, r_2)$ and $s' = (s_0, s_1, s_2)$, follow independent multinomial distributions with indices R and S and respective probabilities $p' = (p_0, p_1, p_2)$ and $q' = (q_0, q_1, q_2)$, where $p_i = P(M = i \mid Y = 1)$ and $P_i = P(M = i \mid Y = 0)$. Alternatively, U can be written as

$$U = \sum_{i=0}^{2} \frac{X_i}{N} [Sr_i - Rs_i]. \tag{1}$$

The variance of U is $V = Var(U) = (1-\phi)^2 RX'\Sigma_p X + \phi^2 SX'\Sigma_q X$, where Σ_p denotes the correlation matrix for the genotype counts for the cases with $(\Sigma_p)_{ii} = p_i(1 - p_i)$ and $(\Sigma_{ik}) = -p_i p_k$, and Σ_q is defined analogously for the controls. Under the null hypothesis, H_0, that $p_i = q_i$, $i = 0, 1, 2$, a valid estimate of V is the pooled variance estimate, obtained by using $\Sigma_p = \Sigma_q = \Sigma$ with estimates $\hat{p} = \hat{q} = n/N$, where $n = (n_0, n_1, n_2)$ is the vector of total counts in Table I. To be explicit,

$$\begin{aligned}\hat{V}_0 &= \widehat{Var}U = N\phi(1 - \phi)X'\hat{\Sigma}X \\ &= \frac{RS}{N}\left(\sum_{i=0}^{2} X_i^2 \frac{n_i}{N}\left(1 - \frac{n_i}{N}\right) - \sum\sum_{i \neq j} X_i X_j \frac{n_i}{N}\frac{n_i}{N}\right).\end{aligned}$$

For an alternative hypothesis, H_1, in which $p_i \neq q_i$ for some $i = 0, 1, 2$, the asymptotic power of the two-sided trend test $|U\hat{V}_0^{-\frac{1}{2}}| > z_{1-\alpha/2}$ can be expressed in terms of $E_{H_1}U/N, \sigma_1^2 = Var(U)$ and the limit of $\hat{V}_0/N$ under H_1, denoted by σ_*^2, as

$$\begin{aligned}&P(|U\hat{V}_0^{-\frac{1}{2}}| > z_{1-\alpha/2}) = 1 - \Phi(z_{1-\alpha/2}\sigma_*/ \\ &\sigma_1 - \sqrt{N}E_{H_1}U/\sigma_1) + \Phi(z_{1-\alpha/2}\sigma_*/ \\ &\sigma_1 - \sqrt{N}E_{H_1}U/\sigma_1),\end{aligned} \tag{2}$$

where Φ stands for the standard normal distribution function and $z_{1-\alpha} = \Phi^{-1}(1 - \alpha)$.

COMPUTATION OF THE MOMENTS OF TEST STATISTIC UNDER H_1

Taking expectations of U under the alternative yields $E_{H_1}U = \sum_{i=0}^{2} \frac{X_i RS}{N}[p_i - q_i]$, and

$$\lim_N E_{H_1}\hat{V}_0/N = \phi(1-\phi)\sum_{i=0}^{2} X_i^2(\phi p_i - (1-\phi)q_i (1-\phi p_i - (1-\phi)q_i) - \phi(1-\phi)\sum\sum_{i\neq j} X_i X_j \times (\phi p_i + (1-\phi)q_i)(\phi p_j + (1-\phi)q_j) \quad (3)$$

The calculation of $p_i = P(M = i \mid Y = 1)$ and $q_i = P(M = i \mid Y = 0)$ depends on the extent of linkage disequilibrium between the marker locus and the disease locus and on the disease penetrance for the disease gene. Assume that the disease locus, like the marker locus, is biallelic, with G denoting the disease-causing allele and g the wild-type allele. Denote the penetrances for the disease genotype by $f_k = P(Y = 1 \mid D = k)$, for $k = 0, 1, 2$, where $D = k$ corresponds to the genotype containing k disease-causing alleles. If the marker has no effect on the probability of disease given D, i.e., $P(Y = 1 \mid D, M) = P(Y = 1 \mid D)$, the probabilities are

$$p_i = P(M = i|Y = 1) = \frac{P(M = i, Y = 1)}{P(Y = 1)} = \frac{\sum_k f_k P(M = i, D = k)}{\sum_k P(D = k)}, i = 0, 1, 2.$$

The calculations for the q_i's for the controls are analogous, with the penetrances f_i replaced by $(1-f_i)$. The joint genotype probabilities, $P(M,D)$, for the two-locus biallelic model can be found as functions of the wild-type allele frequencies p_a and p_g, and the linkage disequilibrium coefficient, $\delta = P(ag) - p_a p_g$. Here $P(ag)$ is the probability that a randomly selected haplotype carries both wild-type alleles. For a homozygote wild-type, for example, the genotype probability is $P(aagg) = (p_a p_g + \delta)^2$. A complete list of genotype probabilities for a two-locus Mendelian model can be found in Khoury et al. [1993, Table 8-5, p. 257]. Note that p_a is the frequency of the wild-type allele at the disease locus, and $1 - P_g$ the frequency of the disease-producing allele, G. In the special case when the marker is the disease gene, equation (3) yields $p_i = f_i P(D = i)/\sum f_k P(D = k)$ and $q_i = (1 - f_i)P(D = i)/\sum (1 - f_k) P(D = k)$, as in Slager and Schaid [2001].

The null hypothesis H_0: $p = q$ is true if either the disease gene imparts no added risk, i.e., $f_0 = f_1 = f_2$, or if the marker is in linkage equilibrium with the disease gene, i.e., $\delta = 0$, in which case $P(M = i, D = k) = P(M = i)P(D = k)$. In both situations, no association or linkage equilibrium, (3) reduces to $p_i = q_i = P(M = i)$.

The power and sample size for the Armitage trend test therefore depend on the allele frequencies at the marker and disease locus, the penetrances for the disease gene, and the amount of linkage disequilibrium. In Results, we present numerical studies and simulations to describe the dependence of the power on those parameters.

SAMPLE SIZES FOR THE SIBLING CASE-CONTROL DESIGN

The Armitage trend test assumes a random sample of cases and controls, all of which are independent of each other. Many designs for association tests are based on family data in which phenotypes are correlated within family. Here, we investigate the power of a marker-based test for the sibling case-control design.

Let $X_{ij}(D)$ denote the score for the jth sibling of the ith sibship associated with the genotype $D = 0$, 1, or 2 based on a biallelic disease gene. We assume that the penetrance function for the disease gene for the jth sibling of the ith sibship follows a logistic model

$$\log it(p_{ij}) \log it P(Y_{ij} = 1|z_{j,} X_{ij}) = \mu + \sigma_z z_i + \beta X_{ij}^D \quad (4)$$

where Y_{ij} denotes the disease status, $X_{ij}^D = X(D_{ij})$ is the score associated with marker genotype D_{ij}, and z_i is a random familial effect for the ith family with $E(z_i)=0$ and $Var(z_i)=1$, which is independent of the genetic effects.

We assume a sample of N discordant sib pairs satisfying $Y_{i\cdot} = \sum_{j=1}^{2} Y_{ij} = 1$. Note that in distinction to the previous section, the total sample size is $2N$. The log-likelihood under model (4), conditional on $Y_i = 1$, for N sibships is

$$L(Y_{1,}, \ldots, Y_N, \beta) = \sum_{i=1}^{N}\left[\beta\sum_{j=1}^{2} Y_{ij}X_{ij}^D - \ln\left\{\sum_{j=1}^{2}\exp(\beta Y_{ij}X_{ij}^D)\right\}\right] \quad (5)$$

and the corresponding score test for $\beta = 0$ is

$$U = \sum_{i=1}^{N} U_i \equiv \sum_{i=1}^{N}\left\{X_{il}^D\left(Y_{i1} - \frac{1}{2}\right) + X_{i2}^D\left(Y_{i2} - \frac{1}{2}\right)\right\}. \quad (6)$$

For the marker, we use (6) with X_{ij}^D replaced by X_{ij}^M as in the unrelated case-control design. Only

siblings with different genotypes contribute to the conditional likelihood (5). Under the null hypothesis $\delta = 0$, and regardless of the extent of linkage between the marker and the disease locus, X_{ij}^M is independent of Y_{ij}, for $j = 1, 2$. Thus the expectation of U is zero when $\delta = 0$. To standardize U, we use the empirical variance estimate

$$\hat{V}_0 = N/(N-1)\sum_{i=1}^{N}(U_i - \bar{U})^2 \qquad (7)$$

where $\bar{U} = U/N$. Note that $\hat{V}_0$ is a valid variance estimate even in the presence of residual correlation between sibs due to linkage (the recombination fraction $\theta < 1/2$) between the marker and the disease locus, as the independent unit is the sibship. In the presence of linkage, a standard model-based variance estimate for the score test that assumes $\theta = 1/2$ can underestimate the true variance, inflating the size of the test. Siegmund et al. [2000] proposed a Wald test with a sandwich variance estimate to avoid this problem when testing for association in sibships. Using calculations in the Appendix, we show how to compute the power of the test $|U\hat{V}_0^{-1/2}| > z_{1-\alpha/2}$ from equation (2) for arbitrary values of δ, θ, and the penetrance parameters.

The quantities U and V can be generalized to sibships with k affected and l unaffecected sibs by considering conditional logistic regression obtained from $k : l$ matching. The corresponding sample size calculations become quite involved, however, as the joint distribution of multiple genotypes and markers is required (see Appendix).

COMMENTS AND EXTENSIONS

Solving equation (2) as a function of N for a chosen power yields the required sample size. We use a very good closed-form approximation for the sample size, derived by discarding the last term in equation (2). This results in

$$N = (z_{1-\alpha/2}\sigma_* + z_{1-\beta}\sigma_1)^2/(E_{H_1}U)^2 \qquad (8)$$

where $1-\beta$ denotes the chosen power. This formula can also be used to approximate the change in sample size needed to accommodate a more stringent α level. If one were interested, for example, in a genome-wide association scan, $\alpha^* = 5 \cdot 10^{-8}$ has been suggested by several authors as an appropriate choice, and the corresponding sample size $N^* = N(z_{1-\alpha^*/2}\sigma_* + z_{1-\beta}\sigma_1)^2/(z_{1-\alpha/2}\sigma_* + z_{1-\beta}\sigma_1)^2 \approx N(z_{1-\alpha^*/2} + z_{1-\beta})^2/(z_{1-\alpha/2} + z_{1-\beta})^2$. For example, with a power of 80%, using $\alpha^* = 5 \cdot 10^{-8}$ instead of $\alpha = 0.05$ requires increasing the sample size by a factor of $(5.33 + 0.84)^2/(1.96 + 0.84)^2 = 4.9$.

Equation (8) can also be used to approximate the dependence of the sample size on δ. If $\sigma^* = \sigma_1$ for small δ, then by using a Taylor expansion around $\delta = 0$, we get

$$N \approx (z_{1-\alpha/2} + z_{1-\beta})^2(\sigma_1^2(0) + \delta\sigma_1^{2'}(0))/(E_{H_1}U(0) + \delta E_{H_1}U(0)')^2.$$

As $E_{H_1}\, U = 0$ if $\delta = 0$, we have

$$N \approx (z_{1-\alpha/2} + z_{1-\beta})^2\sigma_1^2(0) + \sigma_1^2(0))/\delta^2(E_{H_1}U(0)')^2 + O(1/\delta).$$

The sample size thus varies with δ^{-2} for small δ. We assess the accuracy of this approximation in Results.

Because one will not in general know which marker allele, A or a, is in positive linkage disequilibrium with a putative disease allele G, it is desirable to choose a set of scores such that the test statistic is invariant to the choice of which allele is the ''marker allele''. The additive scores $X = [0, 1, 2]$ satisfy this criterion, but dominant scores $X = [0, 1, 1]$ or recessive scores $X = [0, 0, 1]$ do not. For this reason, we shall study the power of the test based on additive scores more extensively, and also consider the sample size that it would require when a dominant or recessive model holds. The additive scores correspond to additivity of the logit of the penetrance. For small penetrances, additivity on the logistic scale corresponds approximately to multiplicative penetrances, $f_0, f_1 = f_0 e^{\beta}, f_2 = f_0 e^{2\beta}$.

RESULTS

PARAMETERS FOR SAMPLE SIZE CALCULATIONS

The amount of linkage disequilibrium, δ, depends on the allele frequencies at both loci in the population. To facilitate comparison between different situations, we follow Tu and Whittemore [1999] and use a standardized version of δ, denoted by D', that was introduced by Lewontin [1964]:

$$D' = \begin{cases} \delta/\min(p_G p_a, p_g p_a), & \delta > 0 \\ \delta/\min(p_g p_A, p_G p_A), & \delta < 0 \end{cases}$$

where the denominator is the absolute maximum δ that could be achieved given the allele frequencies. The values of D' range from –1 to 1. In our simulations and computations, we restrict D' to be

positive. We use level α=0.05 and power 0.80 throughout. We also fix the attributable risk for genetic effects, AR, defined as AR=1 – $f_0/P(Y$=1), and the disease prevalence $P(Y$=1) in the population to determine parameters in the logistic models and p_g.

INDEPENDENT CASES AND CONTROLS

We study the sample sizes required for additive scores when additive, dominant, or recessive models hold (Figs. 1–4), for reasons explained above. The penetrances for the disease locus are modeled using a logistic regression model as $P(Y = 1 \mid D)$=exp(μ+ βX_i)/(1 – exp(μ + βX_i)), where $X_i = X(D = i)$ denotes the score for the disease genotype D. Because we set X_0=0 for all models, fixing AR and $P(Y = 1)$ allows us to obtain μ for the underlying true inheritance model by solving f_0=exp(μ)/(1 – exp(μ))=(1 – AR)$P(Y = 1)$. We study the alternative $\beta = 1$. The allele frequency of the wild-type allele at the disease locus, p_g, is found from solving $P(Y = 1) = f_0 p_g^2 + 2(1 - p_g)p_g f_1 + f_2(1 - p_g)^2$.

Figures 1–4 plot N against $100D'$, the logarithm of the percent of maximum linkage disequilibrium, on the $\log_{10}$ scale for various choices of marker allele frequencies p_a. Figures 1–4 correspond respectively to (AR, $P(Y = 1)$)=(0.3, 0.01), (0.3, 0.1), (0.1, 0.01), and (0.1, 0.1).

Four factors dominate the variation in required sample sizes seen in Figures 1–4: (1) the degree of discordance between p_g and p_a, which determines the maximal linkage disequilibrium, $100D'$; 2) the percentage of maximal linkage disequilibrium; 3) the attributable risk (AR); and 4) the mode of transmission.

Figure 1 illustrates the importance of discordance between p_g and p_a. For each inheritance model, the plots run nearly parallel, and the vertical displacements represent the effects of mismatching allele frequencies. For example, in the dominant model, $p_g = 0.86$ matches $p_a = 0.8$ well, whereas the locus with $p_a = 0.5$ requires

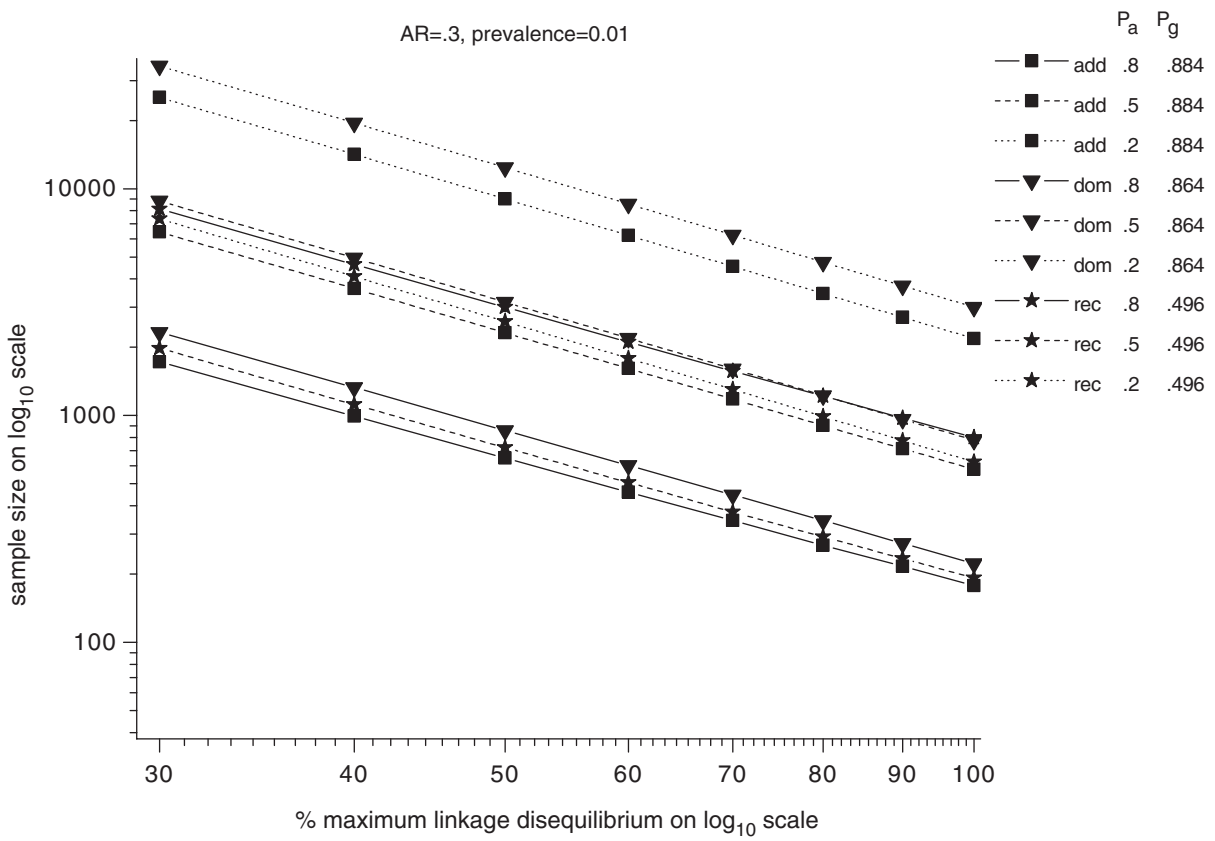

Fig. 1. Required total number of cases and controls (on a $\log_{10}$ scale) plotted against $\log_{10}$ ($100D'$) for various inheritance models and marker allele frequencies. The two-sided 0.05 level test statistic is based on additive scores, and has power of 0.8. Prevalences of disease allele and wild-type penetrance are determined by conditions $AR = 0.3$ and disease prevalence=0.01.

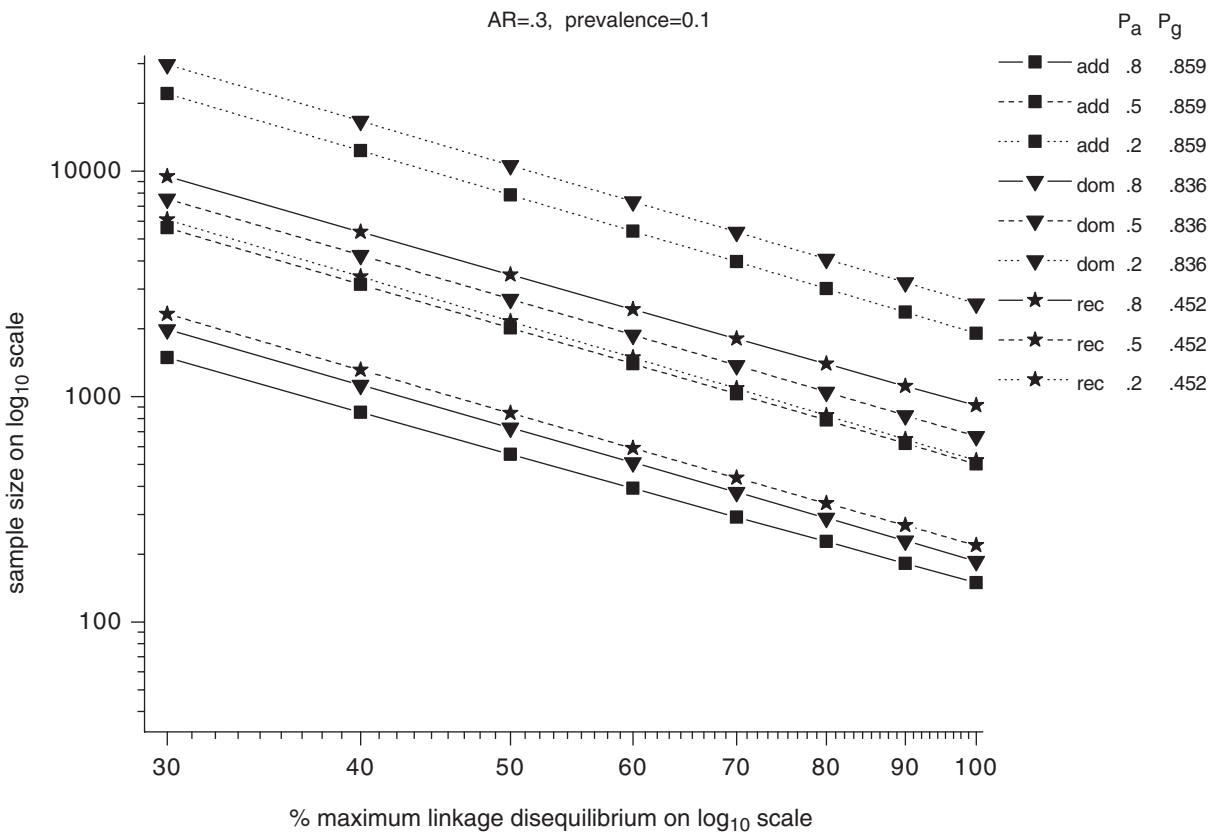

Fig. 2. As in Figure 1, except $AR = 0.3$ and disease prevalence=0.1.

about sixfold larger samples, and the locus with $p_a = 0.2$ requires about tenfold larger samples. The effects of mismatched allele frequencies are similar for the additive model, but not as extreme as for the recessive model. Table II illustrates the effect of mismatched allele frequencies on sample size requirements for the settings in Figures 1–4 for $D' = 1$ by comparison with the sample size required for the disease gene. Even when the allele frequency of the marker is close to that of the disease allele, as in the case of an additive model with $AR = 0.3$, $P(Y = 1)$=0.1, $p_g = 0.8586$, and $p_a = 0.8$, the sample size required for the marker is 37% larger than for the disease locus. In all other situations, the required sample size increases are even bigger.

The effect of the percent maximal linkage disequilibrium is also large (Figs. 1–4). Indeed, the ratio of the required sample size with $100D' = 30\%$, compared to $100D' = 100\%$, is about 10 for all the loci in Figures 1–4.

To assess the approximation that the required sample size varies inversely with δ^2 [see also Risch and Teng, 1998], we calculate the slopes in Figures 1–4. The slopes range from −1.90 to −2.04, in reasonable agreement with –2 given by the approximation.

Comparing Figures 3 and 4 with Figures 1 and 2, one sees that larger samples are required to detect genetic factors with smaller attributable risks, but disease prevalence has relatively little impact. Mode of transmission also affects sample size, especially in the presence of substantial mismatch between p_g and p_a for $AR = 0.3$. For $AR = 0.1$ with substantial mismatch, recessive inheritance requires smaller samples than dominant inheritance, whereas for $AR = 0.3$ and good matching between p_g and p_a, recessive inheritance requires larger samples.

We also calculated required sample sizes using additive, dominant, and recessive scores, respectively, for additive, dominant, and recessive modes of inheritance and under the unrealistic assumption that one knew which marker allele was in positive linkage disequilibrium with the disease allele. These scores would be optimal if

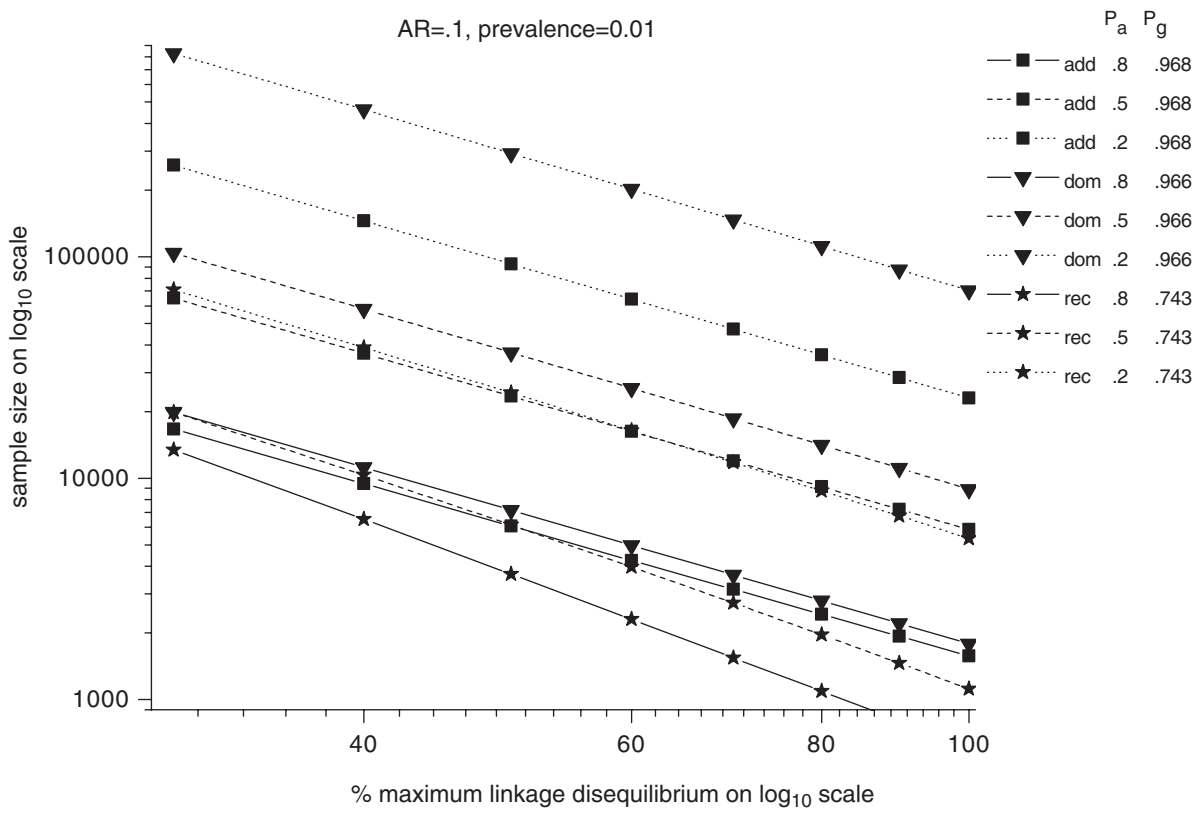

Fig. 3. As in Figure 1, except $AR = 0.1$ and disease prevalence=0.01.

the marker allele was exactly the disease allele, or if $p_a = p_g$ and $D' = 1$. We calculated how much efficiency is lost by using additive scores when, in fact, dominant (recessive) scores were optimal by computing the ratio of sample sizes required with additive scores to that required by dominant (recessive) scores when the inheritance model was dominant (recessive) (Table III). For dominant inheritance, the additive scores only require at most 11% larger samples than dominant scores. This occurs, as expected, when there is substantial linkage disequilibrium. With weak linkage disequilibrium, additive scores for the marker may require 67% smaller sample sizes than dominant scores. Thus scores that are optimal for the disease allele are not necessarily efficient for the marker. Similar comments apply to recessive inheritance, except that additive scores can require up to 44% larger samples with tight linkage disequilibrium, or about 60% smaller samples with weak linkage disequilibrium. Unreported plots for the dominant and recessive scores yield similar qualitative conclusions as in Figures 1–4, except that the slopes are not all near –2.

THE SIBLING CASE-CONTROL DESIGN

The calculations for μ and p_g for the sibling case-control design are more complex, as the model also includes the normally distributed random intercept, z_i, to account for residual correlation in the ith sib pair. In addition to the AR, and disease prevalence, we fixed the value of σ_z^2. We solved the equations $f_0 = (1 - AR)P(Y = 1) = \int \exp(\mu + z)/(1 - \exp(\mu + z))dF(z)$ numerically for μ, using Gaussian quadrature for the integrations, in MATLAB 6.1 [Mathworks, Inc., 1999]. Then we solved for p_g by fixing $P(Y = 1)$, as above. Again, the alternative hypothesis we studied was $\beta = 1$.

The second additional parameter that enters the model for the sibling case-control design is the recombination fraction θ, that lessens the degree of linkage disequilibrium that can be obtained. The sample-size requirements to achieve 80% power increase by about 4% for $\theta = 0.01$ compared to $\theta = 0$ for all choices of linkage disequilibrium and all genetic models in Figures 1–4. For $\theta = 0.1$, on the other hand, the impact on sample size is substantial. For example, for $p_a = 0.8$ and $\sigma_z^2 = 0$,

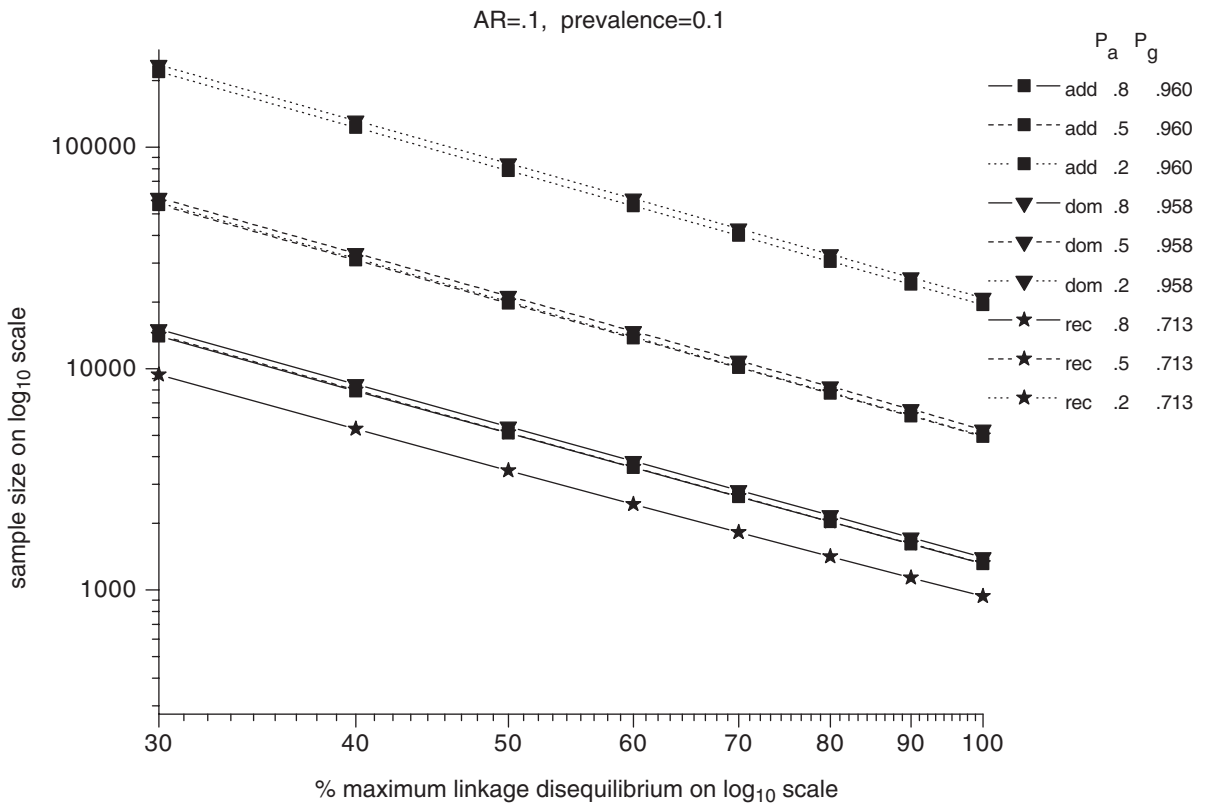

Fig. 4. As in Figure 1, except $AR = 0.1$ and disease prevalence=0.1.

TABLE II. Independent Cases and Controls: Sample-Size Requirements Using Additive Scores for Various Choices of Marker Allele Frequencies for $D' = 1, \alpha = 0.05$, Power=80%

				Sample size for marker with p_a			
AR	$P(Y = 1)$	Model	p_g[a]	0.8	0.5	0.2	Sample size for disease gene
0.1	0.1	Additive	0.9596	1,315	4,938	19,430	311
		Dominant	0.9581	1,402	5,299	20,885	330
		Recessive	0.7136	937	1,321	4,966	593
0.1	0.01	Additive	0.9677	1,574	5,858	22,994	337
		Dominant	0.9663	1,694	6,355	25,002	356
		Recessive	0.7426	885	1,605	5,992	649
0.3	0.1	Additive	0.8586	149	501	1,907	109
		Dominant	0.8364	186	666	2,585	151
		Recessive	0.4519	912	219	520	179
0.3	0.01	Additive	0.8837	178	578	2,180	114
		Dominant	0.8636	222	778	3,001	153
		Recessive	0.4958	800	192	623	189

[a]Note that disease allele frequency is $1 - p_g$.

the sample-size requirements to achieve 80% power increase by 50% for all choices of linkage disequilibrium and all genetic models, compared to $\theta = 0$. However, as substantial linkage disequilibrium rarely extends beyond 100 kilobases, values of θ larger than 0.01, which corresponds to approximately 1,000 kilobases, need not be considered.

Residual correlation has a comparably small effect in this model. When σ_z^2 is changed from zero to one, with $\theta = 0$, the sample sizes required to guarantee 80% power increase by 1–3%, depending on the underlying model and marker allele frequency. Other authors noted a weak effect of residual familial correlation on the power of association tests as well [e.g., Shih and Whittemore, 2002].

TABLE III. Independent Cases and Controls: Ratios of Sample-Size Requirements, Using Additive Scores Divided by Those for Optimal Scores for Disease Gene, With $\alpha = 0.05$, Power=80%

		% maximum linkage disequilibrium							
p_g[1]	p_a	30	40	50	60	70	80	90	100
$AR = 0.3$, $P(Y = 1)$=0.01									
Dominant model									
0.86	0.8	0.954	0.974	0.994	1.013	1.030	1.049	1.066	1.083
	0.5	0.718	0.736	0.754	0.773	0.791	0.810	0.830	0.851
	0.2	0.358	0.367	0.376	0.386	0.396	0.406	0.416	0.427
Recessive model									
0.50	0.8	0.415	0.441	0.468	0.494	0.520	0.546	0.571	0.597
	0.5	0.833	0.892	0.956	1.020	1.093	1.168	1.251	1.333
	0.2	0.945	0.966	0.987	1.010	1.034	1.060	1.087	1.117
$AR = 0.3$, $P(Y = 1)$=0.1									
Dominant model									
0.84	0.8	0.961	0.983	1.006	1.026	1.047	1.066	1.085	1.107
	0.5	0.724	0.743	0.763	0.784	0.804	0.826	0.847	0.871
	0.2	0.361	0.371	0.381	0.392	0.403	0.414	0.425	0.438
Recessive model									
0.45	0.8	0.403	0.427	0.450	0.473	0.496	0.519	0.542	0.564
	0.5	0.808	0.859	0.910	0.966	1.021	1.080	1.145	1.210
	0.2	0.946	0.967	0.989	1.012	1.036	1.062	1.088	1.116
$AR = 0.1$, $P(Y = 1)$=0.01									
Dominant model									
0.97	0.8	0.907	0.913	0.919	0.925	0.931	0.936	0.942	0.947
	0.5	0.680	0.685	0.689	0.694	0.699	0.703	0.708	0.713
	0.2	0.340	0.342	0.344	0.346	0.349	0.351	0.353	0.356
Recessive model									
0.74	0.8	0.637	0.749	0.862	0.978	1.095	1.212	1.330	1.448
	0.5	0.866	0.938	1.013	1.091	1.172	1.256	1.345	1.437
	0.2	0.953	0.975	0.998	1.022	1.046	1.070	1.096	1.122
$AR = 0.1$, $P(Y = 1)$=0.1									
Dominant model									
0.96	0.8	0.909	0.915	0.921	0.928	0.934	0.940	0.946	0.952
	0.5	0.681	0.686	0.691	0.696	0.701	0.706	0.711	0.717
	0.2	0.340	0.343	0.345	0.348	0.350	0.353	0.355	0.358
Recessive model									
0.71	0.8	0.596	0.693	0.793	0.894	0.998	1.103	1.208	1.314
	0.5	0.867	0.940	1.016	1.094	1.176	1.262	1.350	1.442
	0.2	0.953	0.976	0.999	1.022	1.046	1.071	1.096	1.122

[1]Note that disease allele frequency is $1-p_g$.

The sample sizes needed for the sibling case-control design are known to be about twice as large as those needed for the unmatched case-control design for candidate genes [e.g., Witte et al., 1999]. We computed the ratios of sample sizes needed for the sibling case-control design to those needed for the unmatched case-control design with markers for $\sigma_z = 0, \theta = 0$, and various scenarios covered in Figures 1–4. These ratios were based on the use of additive scores, regardless of the actual mode of inheritance. For all situations, the ratios were between 1.89–2.04, and 98% of the ratios were within 3% of 2.0. Thus, when a marker is used instead of a candidate gene, the sibling case-control design requires very nearly twice as many subjects as the unmatched case-control design.

SIMULATIONS

To assess the accuracy of the asymptotic formulas, we simulated 10,000 replicates of outcome data with the calculated sample sizes, independently for various models. For each replicate, the test statistic was calculated, and the true power estimated as the proportion of

replicates which were significant at $\alpha = 0.05$. The empirical power estimates (not shown) indicated excellent agreement with the expected power of 0.80.

DISCUSSION

We evaluated the sample-size requirements for two-sided trend tests with additive scores applied to marker genotypes, both for independent samples of cases and controls and for discordant sib-pairs. We are unaware of previous studies of the power of markers for these important designs.

The main factors influencing sample size, apart from α-levels and relative risk parameters, are: 1) the degree of agreement between the marker allele and disease allele frequencies, which determines the maximal linkage disequilibrium; 2) the percent of maximal linkage disequilibrium present; and 3) the attributable risk from the disease allele. Type of inheritance also plays a role. For a fixed attributable risk, disease prevalence, and in the sibling case-control design, familial aggregation other than that induced by the disease gene of interest, and recombination have a much smaller impact. Note that holding the attributable risk fixed and letting the disease prevalence vary changes the heritability of the disease, i.e., the percentage of total variation explained by the gene. As one reviewer pointed out, had we used a different parameterization and held the recurrence risk, or equivalently, heritability, fixed, we would have seen a strong impact of disease prevalence on power. Previous work studying the transmission disequilibrium test stressed the importance of discrepant allele frequencies and less-than-maximal D' in inflating required sample size [Abel and Müller-Myhsok, 1998; Tu and Whittemore, 1999].

We found that additive scores applied to the marker data can actually be more efficient than dominant scores with dominant inheritance, or recessive scores with recessive inheritance. Thus, the use of additive scores, which do not require knowing which marker allele is in positive linkage disequilibrium with the putative disease allele, may not be very inefficient, and can even be advantageous in some settings.

A sobering implication of our work is that very large sample sizes are required, especially with modest attributable risks like $AR = 0.1$ (Figs. 3, 4). In this circumstance, studies with fewer than 10,000 subjects have adequate power only if marker alleles have frequencies near that of the disease allele and attain at least 50% of the maximal linkage disequilibrium. Figures 3 and 4 also highlight the importance of using a marker whose allele frequency matches that of the putative disease allele frequency. Thus, several markers should be examined in the region of a putative disease locus, to increase the chances of finding one with strong linkage disequilibrium. Required sample sizes for sibling case-control designs are about twice as large as for the unmatched case-control design. For attributable risks of 0.3 (Figs. 1, 2), case-control studies with fewer than 10,000 subjects are required for broader ranges of discrepancy in allele frequencies and less-than maximal linkage disequilibrium.

The large sample-size requirements may partly explain why many genetic associations based on SNPs have not been confirmed in subsequent studies [Hirschhorn et al., 2002]. The probability that a statistically significant association is, in fact, a true association, and not simply the result of statistical noise, depends on the power of the test. This probability is called the positive predictive value [Vecchio, 1966]. For a given prior probability, lower power will result in a lower positive predicitve value, contributing to the chance of false-positive findings.

The two requirements that the marker allele have a frequency close to that of the disease allele and that it attain a substantial fraction of maximal linkage disequilibrium may not be satisfied by randomly selected SNPs within a candidate region. To assure that some of the SNPs have an allele frequency close to that of the putative disease allele, the more common alleles of the selected SNPs should have allele frequencies covering the range 0.5–0.95. An alternative strategy for selecting promising SNPs would be to examine all functional alterations in the same region. In this way, one might hope to find a disease-producing locus within the region, rather than simply a marker in linkage disequilibrium.

If the attributable risk for the locus under study is correctly specified, a second unlinked disease gene will have only minor impact on the required sample size, as indicated by the results for σ_z^2.

Although calculations in this paper focused on a single marker, the calculations have implications for other designs. For example, a whole-genome scan with 10^6 SNPs would require a significance level of $5 \cdot 10^{-8}$ to assure experiment-wise control of significance, based on the Bonferroni inequality. As indicated in Statistical Methods, this would require that the sample sizes be increased by

a factor of 4.9 from those presented in this paper. Likewise, one can adjust the α-level to accommodate screening of markers in a candidate region.

Schork [2002] calculated the power of an allele-based (rather than genotype-based) test for independent cases and controls, using a marker. He used a Bayesian approach to integrate over the distributions of δ and the marker allele frequencies. We obtain similar sample-size requirements for independent cases and controls with additive scores when comparably high penetrances are assumed.

The sibling case-control design and related test statistics were proposed as an alternative to case parent trios for late-onset diseases [e.g., Curtis, 1997], and modified by several authors [e.g., Horvath and Laird, 1998]. Kaplan and Martin [2001] presented power calculations for a class of allele-based association tests in families with genotyped parents and offspring. They assumed complete linkage (θ=0) between the marker and the disease locus, and determined optimal weights for terms involving affected and unaffected offspring. The sibling case-control design with additive scores corresponds to a special case called the weighted sibling transmission equilibrium test (WSTDT). Kaplan and Martin [2001] did not present numerical results on power and sample size, however, nor did they present data on the effect that δ has on those quantities.

Lange and Laird [2002a, b] calculated the power for familiy-based association tests (FBATs) based on disease genes. Our score U_i for the sibling case-control design corresponds to the numerator of an FBAT statistic. Our normalization by an empirical variance estimate (7) differs from the variance estimators studied by Lange and Laird [2002a, b], however, and their power calcuations are therefore not applicable to our statistic for the sibling case-control design, even for the case of a disease gene.

This paper focused exclusively on the evaluation of one SNP at a time. Tests based on genotypes or haplotypes from several SNPs may reduce the sample sizes needed to detect association [e.g., Fallin et al., 2001].

A MATLAB 6.1 program for sample size calculations is available from the first author (R.M.P.) upon request.

ACKNOWLEDGMENTS

We thank Alisa Goldstein, Hongyu Zhao, Daniel Schaid, and the reviewers for helpful comments.

REFERENCES

Abel L, Müller-Myhsok B. 1998. Maximum-likelihood expression of the transmission/disequilibrium test and power considerations. Am J Hum Genet 63:664–667.

Armitage P. 1955. Tests for linear trends in proportions and frequencies. Biometrics 11:375–386.

Curtis D. 1997. Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319–333.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics 55:997–1004.

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. 2001. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 11:143–151.

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. 2002. A comprehensive review of genetic association studies. Genet Med 4:45–61.

Horvath S, Laird NM. 1998. A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 63:1886–1897.

Kaplan NL, Martin ER. 2001. Power calculations for a general class of tests of linkage and association that use nuclear families with affected and una.ected sibs. Theor Popul Biol 60:193–201.

Kaplan N, Morris R. 2001. Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. Genet Epidemiol 20:432–457.

Khoury MJ, Beaty TH, Cohen BH. 1993. Fundamentals of genetic epidemiology. New York: Oxford University Press.

Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144.

Lange C, Laird NM. 2002a. On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. Genet Epidemiol 23:165–180.

Lange C, Laird NM. 2002b. Power calculations for a general class of family-based association tests: dichotomous traits. Am J Hum Genet 71:575–84.

Lewontin RC. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49–67.

Olson JM, Wijsman EM. 1994. Design and sample-size considerations with a disease locus. Am J Hum Genet 55:574–580.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. Science 273:1516–1517.

Risch N, Teng J. 1998. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases—I. DNA pooling. Genome Res 8:1273–1288.

Sasieni PD. 1997. From genotypes to genes: doubling the sample size. Biometrics 53:1253–1261.

Schaid DJ, Rowland C. 1998. Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. Am J Hum Genet 63:1492–1506.

Schork NJ. 2002. Power calculations for genetic association studies using estimated probability distributions. Am J Hum Genet 70:1480–1489.

Shih MC, Whittemore AS. 2002. Tests for genetic association using family data. Genet Epidemiol 22:128–145.

Siegmund KD, Langholz B, Kraft P, Thomas DC. 2000. Testing linkage disequilibrium in sibships. Am J Hum Genet 67:244–248.

Slager SL, Schaid DJ. 2001. Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. Hum Hered 52:149–153.

Mathworks, Inc. 1999. Optimizations Toolbox user's guide.

Thompson EA, Deeb S, Walker D, Motulsky AG. 1988. The detection of linkage disequilibrium between closely linked markers—RFLPS at the AI–CIII apolopoprotein genes. Am J Hum Genet 42:113–124.

Tu IP, Whittemore AS. 1999. Power of association and linkage tests when the disease alleles are unobserved. Am J Hum Genet 64:641–649.

Vecchio TJ. 1966. Predictive value of a single diagnostic test in an unselected population. N Engl J Med 274:1171–1173.

Weiss KM, Terwilliger JD. 2000. How many diseases does it take to map a gene with SNPs? Nat Genet 26:151–157.

Witte JS, Gauderman WJ, Thomas DC. 1999. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am J Epidemiol 149:693–705.

APPENDIX

COMPUTATION OF MOMENTS OF SIBLING CASE-CONTROL TEST STATISTIC UNDER H_1

To find the power of the test statistic, we need to evaluate the expected value and the variance of U under both H_0 and the alternative H_1 conditionally on $Y=1$. Noting that the conditional expectation can be rewritten iteratively as $E_{Y,\mathbf{X}^M|Y=1}U = E_{\mathbf{X}^M}E_{Y,Y=1|\mathbf{X}^M}U/P(Y.=1)$, where $\mathbf{X}^M = (X_1^M, X_2^M)$, and $\mathbf{X}^D = (X_1^D, X_2^D)$, we get

$$\mu = E_{\mathbf{X}^M}(X_1^M - X_2^M)\{P(Y_1=1, Y_2=0|\mathbf{X}^M) - P(Y_1=0, Y_2=1|\mathbf{X}^M)\}/2P(Y=1). \quad (9)$$

Under the null hypothesis that X^M is independent of Y, $P(Y_1, Y_2|\mathbf{X}^M) = P(Y_1, Y_2)$. Thus the term in curly brackets in (9) can be pulled out of the expectation, which results in $E(X_1^M - X_2^M) = 0$, and thus $EU = 0$. Also note that under the model (4) with $\beta = 0$, the term in curly brackets vanishes by exchangeability of Y_1 and Y_2. The denominator is given by

$$P(Y=1) = \sum_{\mathbf{X}^D}\{P(Y_1=1, Y_2=0|\mathbf{X}^D) + P(Y_1=0, Y_2=1|\mathbf{X}^D)\}P(\mathbf{X}^D).$$

The second moment of the score statistic is computed likewise from

$$\mu_2 = \frac{1}{4P(Y=1)}E_{\mathbf{X}^M}\{(X_1^M - X_2^M)^2[P(Y_1=1, Y_2=0|\mathbf{X}^M) + P(Y_1=0, Y_2=1|\mathbf{X}^M)]\} \quad (10)$$

which under H_0 reduces to $\mu_2 = \frac{1}{4}E(X_1^M - X_2^M)^2$. We use the empirical variance of the scores in equation (7) to estimate the variance of U. Under $H_1, \hat{V}_0/N$ converges to $\sigma_1^2 = \mu_2 - \mu^2$. Hence, the power of the test $|U\hat{V}_0^{-1/2}| > z_{1-\alpha/2}$ can be computed from equation (2) with $\sigma_* = \sigma_1$.

We obtain the joint penetrance for the siblings conditional on the marker loci from the assumed conditional independence of marker and disease status, given the disease gene and the random familial effect a, as

$$P(Y_1, Y_2|\mathbf{X}^M) = \sum_{\mathbf{X}^g}\int P(Y_1|z, X_1^D)P(Y_2|z, X_2^D) \times dF(z)P(\mathbf{X}^D|\mathbf{X}^M) \quad (11)$$

where F denotes the distribution function of the random familial effects. The terms $P(Y|z, X^D)$ are computed from the logistic model (4).

The joint probabilities $P(\mathbf{X}^D, \mathbf{X}^M)$, and hence $P(\mathbf{X}^D|\mathbf{X}^M)$ in (11), can be computed from the probabilities of the joint genotypes $P(M_1, D_1, M_2, D_2)$ of siblings one and two, because the scores are known functions of the genotypes. This computation is described next.

COMPUTATION OF JOINT GENOTYPE PROBABILITIES FOR TWO SIBLINGS

For ease of exposition, let $\mathbf{g}_i = (M_i, D_i), i = 1, 2$, denote the genotype at the marker and disease locus for sibling i and let $\mathbf{h}_i = (h_{im}, h_{ip})$ denote the haplotype pair of sibling i; $\mathbf{h}_i$ is comprised of the maternal haplotype h_{im} and the paternal haplotype h_{ip}. Define the maternal haplotype pair $\mathbf{h}_m = (h_m^1, h_m^2)$ and the paternal haplotype pair $\mathbf{h}_p = (h_p^1, h_p^2)$. Using conditional independence of the haplotypes of the offspring given the parental haplotypes, and independence of the parental haplotypes, we compute the joint probabilities of haplo-types of the siblings as

$$P(\mathbf{h}_1, \mathbf{h}_2) = \sum_{\mathbf{h}_m, \mathbf{h}_p} P(\mathbf{h}_1, \mathbf{h}_2|(\mathbf{h}_m, \mathbf{h}_p)P(\mathbf{h}_m, \mathbf{h}_p) = \sum_{\mathbf{h}_m, \mathbf{h}_p} P(\mathbf{h}_1, |\mathbf{h}_m, \mathbf{h}_p)P(\mathbf{h}_2|\mathbf{h}_m, \mathbf{h}_p)P(\mathbf{h}_m)P(\mathbf{h}_p). \quad (12)$$

The transmission function $P(\mathbf{h}_i|\mathbf{h}_m, \mathbf{h}_p) = P(h_{im}|\mathbf{h}_m)P(h_{ip}|\mathbf{h}_p)$ gives the probability that a mother with haplotype pair $\mathbf{h}_m$ and a father with haplotype pair $\mathbf{h}_P$ produce an offspring with haplotype pair $(h_{im'}|h_{ip})$. The gametic transmission probabilities $P(h_{im}|\mathbf{h}_m)$ and $P(h_{ip}|\mathbf{h}_p)$ are simple functions of the factor 1/2 and the recombination fraction for the two loci, θ, and its complement, $1 - \theta$.

We assume that the probability of each parental haplotype is determined by the allele frequency

and the linkage disequilibrium coefficient. For example, $P(h_m^1 = \{ag\}) = P(g)P(a) - \delta$.

The joint sibling haplotype probabilities computed from (12) define the joint sibling genotype probabilities $P(\mathbf{g}_1, \mathbf{g}_2)$. In fact, except for terms involving doubly heterozygous genotypes $P(\mathbf{g}_1, \mathbf{g}_2) = P(\mathbf{h}_1, \mathbf{h}_2)$ i.e., the genotype probabilities are equal to the corresponding haplotype probabilities. If the joint genotypes for one sibling are doubly heterozygous, the joint genotype probability is obtained by summing over the corresponding joint two-haplotype probabilities. If the joint genotypes of both siblings are doubly heterozygous, the joint genotype probability is obtained by summing over the four joint haplotype probabilities.